Assembly of complete diploid phased chromosomes from draft genome sequences

Andrea Minio[1], **Noé Cochetel**[1], **Amanda Vondras**[1], **Mélanie Massonnet**[1], **and Dario Cantu**[*, 1]

[1]Department of Viticulture and Enology, University of California Davis, Davis, CA 95616

De novo **genome assembly is essential for genomic research. High-quality genomes assembled into phased pseudomolecules are challenging to produce and often contain assembly errors caused by repeats, heterozygosity, or the chosen assembly strategy. Although algorithms exist that produce partially phased assemblies, haploid draft assemblies that may lack biological information remain favored because they are easier to generate and use. We developed HaploSync, a suite of tools that produces fully phased, chromosome-scale diploid genome assemblies and performs extensive quality control to limit assembly artifacts. HaploSync uses a genetic map and/or the genome of a closely related species to guide the scaffolding of a diploid assembly into phased pseudomolecules for each chromosome. It compares alternative haplotypes to identify and correct misassemblies independent of a reference, fills assembly gaps with unplaced sequences, and resolves collapsed homozygous regions. In a series of plant, fungal, and animal kingdom case studies, we demonstrate that HaploSync increases the assembly contiguity of phased chromosomes, improves completeness by filling gaps, corrects scaffolding, and correctly phases highly heterozygous, complex regions.**

haplotype phasing | diploid genomes| assembly error correction | hybrid genome assembly | chromosome anchoring

Correspondence: ***dacantu@ucdavis.edu***

Introduction

Affordable high-throughput DNA sequencing and novel assembly tools have made high-quality genome assemblies and genome research attainable and abundant. Long-read DNA sequencing technologies, like those developed by Oxford Nanopore (Oxford Nanopore Technologies Ltd., Oxford Science Park, UK) and Pacific Biosciences (Pacific Biosciences Inc, Menlo Park, California, USA), are now the preferred methods for reference genome sequencing. The assemblies produced using these technologies are more contiguous and complete than assemblies constructed using short reads and better represent repetitive content (1–3). Hybrid approaches that integrate optical and linkage maps, chromosome conformation capture, or artificial chromosomes (4–9) can achieve assemblies of complete, full-length chromosomes (10–12).

Partially phased diploid assemblies have revealed astounding and previously unappreciated genomic complexity (2, 13–22). Despite progress, high heterozygosity and repetitive content continue to impair assembly, often inflating primary assembly size (17, 23–26) and undermining the ability to phase diploid assemblies, an important advantage of long read sequencing. A common strategy used to cope with genome complexity is to assemble a haploid representation of a diploid genome. This can be achieved by including a consensus sequence at heterozygous sites or by including only one allele's sequence (27–32). Both choices discard information (20).

Available assembly strategies treat diploid information as if it has little utility (23, 33–37). Some current linkage map-based chromosome reconstruction and reference-guided scaffolding tools consider each haplotype separately and ignore the relationship between them, rather than use that information to attempt recovering unplaced sequences and fill gaps (38, 39). Though quality control is an integral part of the assembly procedure, the relationship between haplotypes is never included in quality control processes. As a result, the assembly depends entirely on an assessment of sequence fragmentation and similarity to a related species' genome (40).

Here, we present HaploSync, an open-source package that scaffolds, refines, and fully phases diploid and chromosome-anchored genomes. HaploSync leverages the relationship between haplotypes to improve the quality and accuracy of assemblies, separate haplotypes while reconstructing pseudomolecule sequences, and recover a location for genomic regions that cannot be placed during other assembly steps. Quality controls are implemented at each step to check for and correct assembly errors. HaploSync was benchmarked using five diploid species with different levels of heterozygosity and from the plant, animal, and fungal kingdoms. For each species, HaploSync delivered a completely phased, chromosome-scaled genome with a quality comparable to the assemblies considered as references for each species. HaploSync, its manual, and tutorials for its use are freely available at `https://github.com/andreaminio/haplosync`.

Materials and methods

HaploSync has six modules: HaploSplit, HaploDup, HaploBreak, HaploFill, HaploMake, and HaploMap. The overall HaploSync workflow is summarized in Figure 1. HaploSync accepts draft genome sequences or assembled pseudomolecules as input, preferably with minimally collapsed heterozygosity and no haploid consensus sequences. Allele phasing is unnecessary *a priori*. The tool is applicable to conventional haploid and diploid-aware assemblies.

HaploSplit. HaploSplit uses external information to associate draft assembly sequences with original chromosomes, then sorts and orients them in pseudomolecules using directed adjacency networks. External information can be a

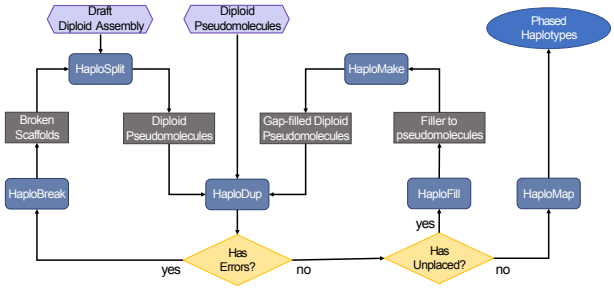

Fig. 1. The HaploSync pipeline builds and refines haploid and diploid genome assemblies. The diploid-aware pipeline can deliver fully phased diploid pseudomolecules using a draft diploid assembly or diploid pseudomolecules as input.

map of sorted and unique genetic markers (Figure 2) and/or the genome assembly of a closely related species (Supplemental Figure 1). Quality control to identify misassemblies in the input draft sequences is possible when markers are available. HaploSplit's performance depends on the quality of information used to anchor sequences. Genetic maps are preferred to a reference genome sequence because they are not biased by sequence homology or structural variability. Marker density, distribution evenness, and order precision determine the efficiency of the map. Ideally, sequences with at least one marker are placed in a pseudomolecule. HaploSync is able to detect and ignore noise, but shortages of usable markers and unmatched sorting order prevent sequence placement. Briefly, each draft sequence is assigned to a genomic region and oriented using reliable markers in the genetic map or colinearity with the guide genome. HaploSplit considers markers reliable if a maximum of two copies are present in all sequences and sorted coherently within the linkage group. Then, a directed adjacency network is created by connecting non-overlapping sequences along a map or guide sequence. Finally, the two tiling paths that maximize the number of covered markers or the number of bases shared with the guide sequence are selected for each chromosome to generate pseudomolecule scaffolds.

HaploDup. HaploDup (Figure 3 and Supplemental Figure 2) exerts diploid-aware quality control over pseudomolecule sequences, allowing the user to identify misassemblies and expose conflicts that prevent proper sequence placement. HaploDup generates multiple sets of interactive plots to accomplish this. By comparing pseudomolecule sequence directly and integrating structural (contig and scaffold composition) and feature (e.g., markers and genes)information, the user can identify the origins of misassemblies that need editing (Figure 3A). Unplaced sequences are also compared with the pseudomolecules and, by visuazing the relationship between them (e.g., matching marker usage), allows to identify and resolve conflicts that prevent sequence placement into the expected linkage groups (Figure 3B).

HaploBreak. HaploBreak (Supplemental Figure 3) automatically searches for and breaks scaffolds at sequencing gaps (i.e stretches of "N" characters between contigs) or at the nearest known junction.

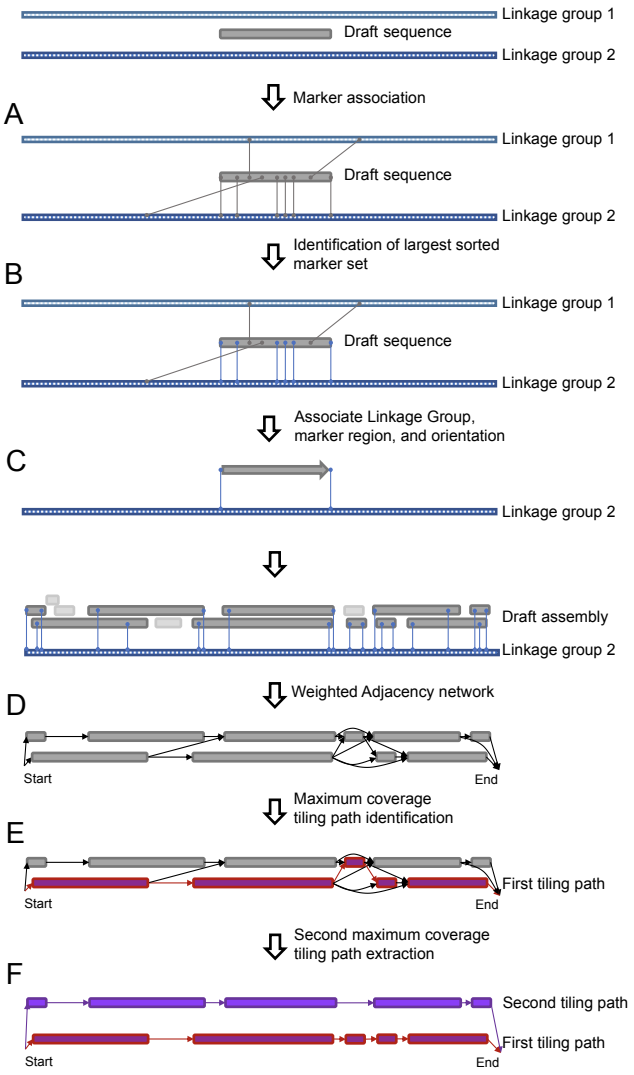

Fig. 2. The HaploSplit procedure using genetic markers as input. A) The procedure identifies marker positions in the draft sequences. B) The longest sorted set of markers is identified for each draft sequence. C) Each sequence is assigned to a unique genomic region in the map (linkage group) and oriented. D) A directed adjacency network of non-overlapping sequences is built for each linkage group. E) The tiling path that maximizes the number of covered markers is selected for the first haplotype. F) Sequences belonging to the first haplotype are removed from the adjacency network and the second-best tiling path is used to scaffold the second haplotype.

HaploFill. HaploFill (Supplemental Figure 4) uses the relationship between homologous pseudomolecule scaffolds to improve the assembly's completeness by integrating unplaced sequences where scaffolding gaps occur. Gaps can have two technical origins: i) a gap in sequence will occur when there is insufficient reliable information to fill the region (e.g., lack of digestion sites in optical maps, lack of markers in HaploSplit, multiple alternative sequences might be used); ii) a gap in in sequence will occur if there is a shortage of sequence available for placement (e.g., homozygous regions were assembled as a single copy).

To find suitable filler, the two haplotypes of each chromosome are compared and supporting information for gap fill-

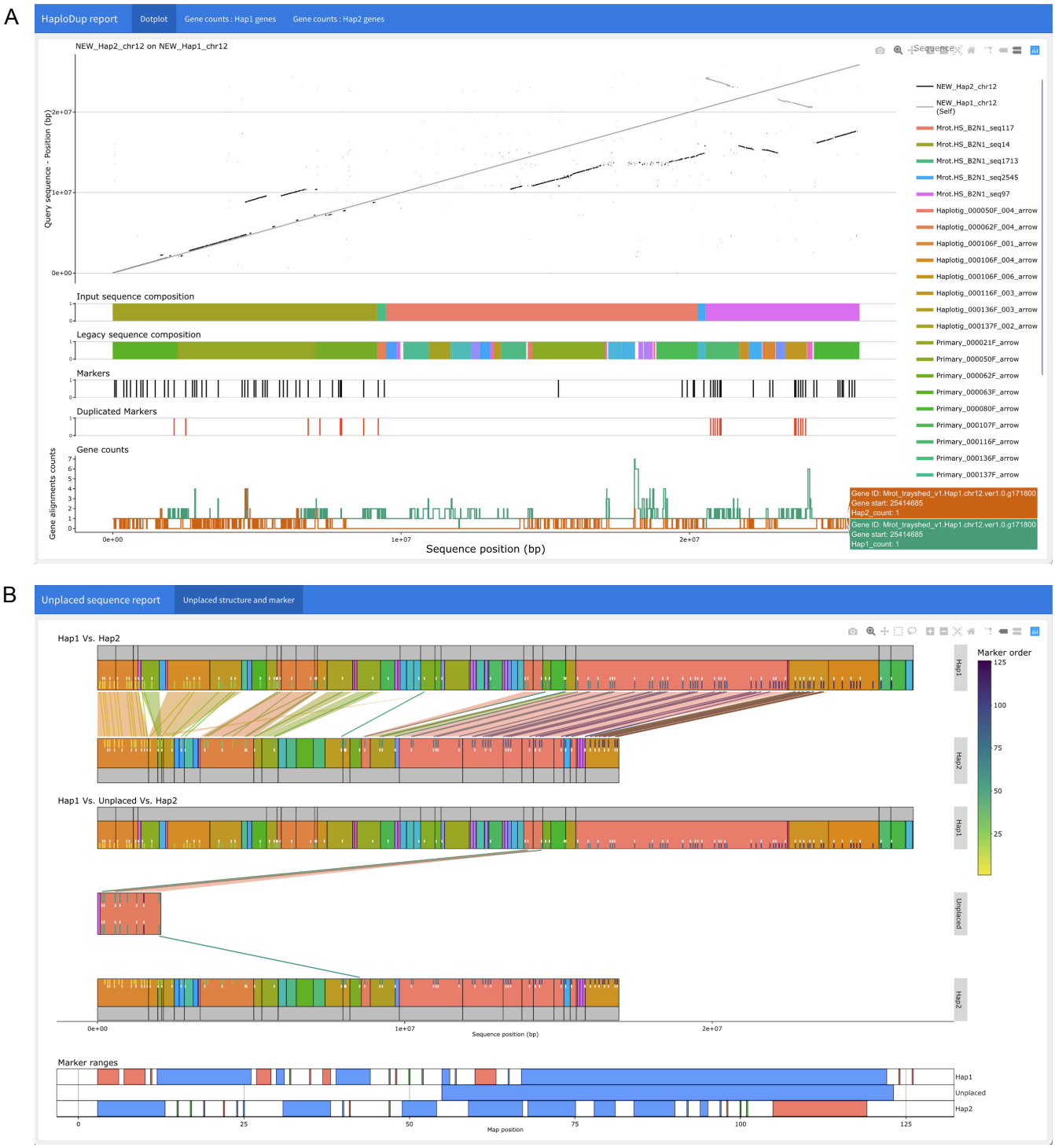

Fig. 3. Example of HaploDup's interactive reports. A) Misassembly quality control. Haplotype 1 and Haplotype 2 sequences are compared in a dotplot reporting contig and scaffold composition. Marker positions and gene copy number across haplotypes are also reported. Within-haplotype duplications caused by fusing primary and alternative alleles can be easily identified. In this example, hybrid scaffolding errors inherited by HaploSplit are visible on *M. rotundifolia* chromosome 12 pseudomolecules. Both alleles and, a consequence, both copies of the markers were placed in the same sequence. B) Unplaced sequence quality control. Marker content between pseudomolecules is compared to evaluate conditions that might prevent the inclusion of unplaced sequences. Color-coding is used for better contextualization, markers color follow the order on the map, sequences color is used to report known sequences relationships (e.g., primary and its associated haplotigs). In the example, the presence of a marker (dark violet) far from the expected position on the map generates the incompatiblity preventing the sequence placement.

ing is sought in the alternative haplotype. Repeat content and ploidy status are used to avoid filling heterozygous deletions. Unplaced sequences are aligned on supporting regions and the most compatible sequence is assigned to fill the gap. If no good candidate can be retrieved and the region is diploid and non-repetitive, homozygous filler is generated from the alternative allele.

HaploMake. HaploMake automates the construction of new assemblies from structure files (e.g., AGP files, and HaploS-

plit and HaploFill outputs). It generates new sequences and updates an existing annotation with new coordinates. The ends of adjacent regions in the structure files can be checked for overlaps with Nucmer (41). The coordinates of adjacent regions can be corrected by adjusting junction positions. This avoids duplicating genomic content in the final sequence and can be done without altering the gene annotation (Supplemental Figure 5).

HaploMap. HaploMap (Supplemental Figure 6) does a pairwise comparison between haplotypes and delivers a map of collinear regions. Local alignments between each pair of sequences are performed with Minimap2 (42) or Nucmer (41). Hits are analyzed to identify a bidirectional tiling path that maximizes the identity between the two sequences. The coordinates of the collinear regions forming the bidirectional tiling path are listed in a pairwise, phased map of matching sequences.

Testing datasets. HaploSync performance was tested using a wide range of species and assembly protocols (Table 1). The diploid *Candida albicans* draft assembly (43), built using PacBio reads and FalconUnzip (23), was anchored to chromosomes using the genetic map generated by (44). A diploid genome assembly of *Arabidopsis thaliana* Columbia-0 (Col-0) X Cape Verde Islands (Cvi-0) (23) was anchored using a genetic map from (45). The *Bos taurus* Angus x Brahma genome from (46) was assembled using FalconUnzip, anchored to chromosomes using the genetic map from (47), and integrated with sex chromosome information from the Integrated Bovine Map from Btau_4.0 release available from https://www.hgsc.bcm.edu/other-mammals/bovine-genome-project. HaploFill was applied once to each of these three genomes. The *Vitis vinifera* ssp. *vinifera* cv. Cabernet Franc FPS clone 04 genome was assembled and scaffolded with PacBio reads and Dovetail HiC data (48). *Muscadinia rotundifolia* cv. Trayshed contigs were assembled with FalconUnzip in hybrid scaffolds that used BioNano NGM maps (49). A *Vitis* consensus genetic map (50) was used to anchor both genomes to chromosomes in HaploSplit and followed by several iterations of HaploFill.

Results and discussion

To evaluate HaploSync's performance, five diploid species from three different kingdoms were selected. An F1 progeny of the model plant *Arabidopsis thaliana* (Col-0 x Cvi-0) (23), *Muscadinia rotundifolia* and *Vitis vinifera* for the plants, the bull *B. taurus* Angus x Brahma (46, 47) for the animals and the pathogenic yeast *Candida albicans*. These species are diverse and vary in genome size, chromosome number, repeat content, and amount of heterozygosity. Long sequencing reads, genetic maps, and public reference genomes are available for those species. Together, these are ideal samples with which to test HaploSync.

HaploSync performance across species. HaploSync produced high-quality genomes for all five species (Table 1). The resulting assemblies were twice the size of their original haploid assemblies, with 1.87 to 2.03 times the gene space represented (Supplemental table 1). This indicates that both complete haplotypes were assembled separately. High-density genetic maps and highly contiguous draft assemblies enabled HaploSplit to produce high-quality pseudomolecules that differed 5.8 - 17.8% from expected chromosome sizes. In one iteration, HaploFill increased assembly completeness and reduced the difference in length between haplotypes.

For *C. albicans*, the limited number of markers were used to anchor 2.4 Mbp of sequences to pseudomolecules in HaploSplit. The assembly had the highest share of unplaced sequences (5.2Mbp), but HaploFill was able to recover 17.9% of the missing genomic content in one iteration. The final pseumolecules were up to 97.9% complete.

With 18.6 ± 0.6 markers/Mbp, the genetic map of *B. taurus* autosomal chromosomes was the most dense. HaploSplit produced pseudomolecules almost identical in size relative to the ARS-UCD1.2 genome assembly (51), with Haplotype 1 pseudomolecules deviating by 0.7 ± 0.7% and Haplotype 2 by 6.7 ± 3.0% (Supplemental Table 2, Supplemental Figure 7). HaploFill inserted 151 Mbp, mostly in Haplotype 2 pseudomolecules, reducing missing information in Haplotype 2 pseudomolecules to 1.4 ± 1.9% of ARS-UCD1.2 chromosome sizes. For sexual chromosomes, only a genetic map of the X chromosome with low marker density was available (2.1 markers/Mbp, assembly ver. Btau_4.0 available at https://www.hgsc.bcm.edu/other-mammals/bovine-genome-project). As a consequence, HaploSplit's performance dropped. HaploSplit retrieved 79.8% of the expected 139 Mbp X chromosome. However, HaploFill reduced missing information to 9.7% (Supplemental table 2, Supplemental figure 7). Without markers available, the length of the Y chromosome was only 11% of its expected size (4.5 Mbp).

In plants the high level of polymorphism and structural variation between haplotypes makes assembly and phasing challenging (23).

The high level of heterozygosity in the *A. thaliana* accession used to test HaploSync is caused by sequence variation between its parents, Col-0 and Cvi-0. This led to a primary assembly 17% longer and haplotigs 11.8% shorter (23) than the haploid reference genome. After Haplosync, the two sets of pseudomolecules differed by 3.6% and 6.3% from the haploid reference genome size. This supports the tool's ability to phase duplicated primary content between haplotypes.

Vitis species can be 12% heterozygous (52). Assemblies of the species can exhibit extensive loss of phase between primary sequences and associated haplotigs (17, 23–25, 53). In Cabernet Franc, for example, the primary assembly is inflated by 18.8% and haplotigs are 40.7% shorter than the expected haploid genome size. HaploSync was able to overcome these limitations for both species and placed over 93.0% of the sequences in phased pseudomolecules that were no more than 9.8% different in size. HaploSplit also automatically placed and correctly phased the grape sex determining region (20) in *Muscadinia* and *Vitis* species.

HaploSync performance across different assembly procedures. HaploSync was applied to two grapes, *M. rotundifolia* cv. Trayshed (49) and *V. vinifera* cv. Cabernet Franc (48), to assess its adaptability to genomes assembled using different protocols. Although contigs were produced with PacBio data and FalconUnzip for both draft assemblies, Trayshed and Cabernet Franc were scaffolded with different technologies.

M. rotundifolia underwent hybrid scaffolding with PacBio and a NGM map, which is based on matching optical fingerprints of DNA molecules with *in silico* digestion of assembled sequences. Consequently, large gaps were introduced due to a low density of digestion sites. Systematic errors were observed at highly repetitive and heterozygous regions. Such error was observed at the RUN1/RPV1 locus on chromosome 12 (Supplemental Figure 8). The differential expansion of TIR-NBS-LRR genes between haplotypes (49) may have caused their fusion in the same scaffold. These issues affected 50 hybrid scaffolds (326.2Mbp) and required correction. For Cabernet Franc, scaffolding with HiC data produced chimeric scaffolds due to the presence of diploid information in the primary assembly. There were 108 of scaffolds (449Mbp) for which the both haplotypes were included in the same assembled sequence (Supplemental Figure 9).

After scaffold correction, both genome assemblies were anchored to chromosomes using a *Vitis* consensus genetic map (50). The low marker density (3.5 markers/Mpb) affected the construction of pseudomolecules by HaploSplit and negatively affected HaploSync's performance. Cabernet Franc was most affected, with only 350.8 Mbp and 263.4 Mbp placed on Haplotype 1 and Haplotype 2, respectively (i.e. 75% and 55% of the reference haploid genome). Unpleaceable sequences were nearly half of Cabernet Franc's expected haploid genome size (240Mbp). Trayshed's Haplotype 1 and Haplotype 2 assemblies were more complete and were 374.3 Mbp and 338.8 Mbp long, respectively.

Three iterations of HaploFill were performed on Cabernet Franc's assembly. Each iteration reduced the unplaced sequences nearly by half (Supplemental Figure 10). The final Cabernet Franc pseudomolecules were 456 Mbp for Haplotype 1 and a 411 Mbp for Haplotype 2. There remains 47 Mbp (5.4%) of sequence unplaced. In contrast, only two iterations of HaploFill left 8% of Trayshed sequences unplaced. Haplotype 1 and Haplotype 2 of Trayshed's pseudomolecules were 400 Mbp and 370 Mbp, respectively. The total size of each haplotype in both chromosome-scale assemblies were similar to their expected haploid reference genome (54, 55) and to Cabernet Sauvignon's haplotypes (459 Mbp and 449 Mbp, respectively) (20).

These results emphasize the importance of controlling and correcting the sequences used as input to HaploSplit to prevent scaffolding errors. Although quality and marker density of the genetic map affect the construction of pseudomolecules by HaploSplit, HaploFill generated phased assemblies with few unplaced sequences and sizes similar to their haploid reference genomes.

Sequencing technologies and assembly tools are continuously improving. HaploSync delivers assemblies with unprecedented quality and contiguity that can provide novel insight into genome structure and organization. The HaploSync suite of tools can be used to address some of the remaining impediments to genome reconstruction and improve assembly quality by taking advantage of diploid information that is readily available. HaploSync correctly and completely phases diploid genomes, reconstructs pseudomolecules by recovering missing information, and exerts quality control over the results.

Web resources

HaploSync is freely available for download at GitHub https://github.com/andreaminio/haplosync. Instructions for installation, a full list of dependencies, a description of each tool, and tutorials are available in HaploSync's manual (https://github.com/andreaminio/HaploSync/tree/master/manual).

Data availability

The data used in this study are summarized in Supplemental table 3. Pseudomolecule reconstructions of *Candida albicans* NCYC4145 (43), *Arabidopsis thaliana* Columbia-0 (Col-0) X Cape Verde Islands (Cvi-0) (23), and *Bos taurus* Angus x Brahma (46) are available at Zenodo (https://zenodo.org/record/3987518). *Vitis vinifera* cv. Cabernet Franc cl. 04 (48) and *Muscadinia rotundifolia* cv. Trayshed (49) pseudomolecule assemblies are available at www.grapegenomics.com.

Funding

This work was funded by the National Science Foundation (NSF) award #1741627 and the US Department of Agriculture (USDA)-National Institute of Food and Agriculture (NIFA) Specialty Crop Research Initiative award #2017-51181-26829. It was also partially supported by funds from E.J. Gallo Winery and the Louis P. Martini Endowment in Viticulture.

Authors' contributions

AM and DC conceptualized the project. AM developed the methodology and software. AM and NC performed bioinformatic analyses and tested the software. AM, NC, and AMV wrote the software manual and pipeline walk-through. AM, AMV, MM, and DC wrote the manuscript. DC secured the funding and supervised the project. All authors have read and approved the manuscript.

1. Anthony Rhoads and Kin Fai Au. PacBio Sequencing and Its Applications. *Genomics, Proteomics and Bioinformatics*, 13(5):278–289, 2015. ISSN 22103244. doi: 10.1016/j.gpb.2015.08.002.
2. Philipp Bongartz and Siegfried Schloissnig. Deep repeat resolution—the assembly of the Drosophila Histone Complex. *Nucleic Acids Research*, 47(3):e18–e18, feb 2019. ISSN 0305-1048. doi: 10.1093/nar/gky1194.
3. Pirita Paajanen, George Kettleborough, Elena López-Girona, Michael Giolai, Darren Heavens, David Baker, Ashleigh Lister, Fiorella Cugliandolo, Gail Wilde, Ingo Hein, Iain MacAulay, Glenn J. Bryan, and Matthew D. Clark. A critical comparison of technologies for a plant genome sequencing project. *GigaScience*, 8(3):1–10, 2019. ISSN 2047217X. doi: 10.1093/gigascience/giy163.

4. Charles A. Tilford, Tomoko Kuroda-Kawaguchi, Helen Skaletsky, Steve Rozen, Laura G. Brown, Michael Rosenberg, John D. McPherson, Kristine Wylie, Mandeep Sekhon, Tamara A. Kucaba, Robert H. Waterston, and David C. Page. A physical map of the human Y chromosome. *Nature*, 409(6822):943–945, feb 2001. ISSN 0028-0836. doi: 10.1038/35057170.
5. Jetty S.S. Ammiraju, Yeisoo Yu, Meizhong Luo, Dave Kudrna, Hye Ran Kim, Jose L. Goicoechea, Yuichi Katayose, Takashi Matsumoto, Jianzhong Wu, Takuji Sasaki, and Rod A. Wing. Random sheared fosmid library as a new genomic tool to accelerate complete finishing of rice (Oryza sativa spp. Nipponbare) genome sequence: Sequencing of gap-specific fosmid clones uncovers new euchromatic portions of the genome. *Theoretical and Applied Genetics*, 111(8):1596–1607, 2005. ISSN 00405752. doi: 10.1007/s00122-005-0091-3.
6. Charlotte G. Cole, Owen T. McCann, John E. Collins, Karen Oliver, David Willey, Susan M. Gribble, Fengtang Yang, Karen McLaren, Jane Rogers, Zemin Ning, David M. Beare, and Ian Dunham. Finishing the finished human chromosome 22 sequence. *Genome Biology*, 9 (5):1–11, 2008. ISSN 14747596. doi: 10.1186/gb-2008-9-5-r78.
7. Harris A. Lewin, Denis M. Larkin, Joan Pontius, and Stephen J. O'Brien. Every genome sequence needs a good map. *Genome Research*, 19(11):1925–1928, 2009. ISSN 10889051. doi: 10.1101/gr.094557.109.
8. Martin Mascher, Gary J. Muehlbauer, Daniel S. Rokhsar, Jarrod Chapman, Jeremy Schmutz, Kerrie Barry, María Muñoz-Amatriaín, Timothy J. Close, Roger P. Wise, Alan H. Schulman, Axel Himmelbach, Klaus F.X. Mayer, Uwe Scholz, Jesse A. Poland, Nils Stein, and Robbie Waugh. Anchoring and ordering NGS contig assemblies by population sequencing (POPSEQ). *Plant Journal*, 76(4):718–727, nov 2013. ISSN 09607412. doi: 10.1111/tpj.12319.
9. Janna L. Fierst. Using linkage maps to correct and scaffold de novo genome assemblies: methods, challenges, and computational tools . *Frontiers in Genetics*, 6:220, jun 2015. ISSN 1664-8021. doi: 10.3389/fgene.2015.00220.
10. Shiguo Zhou, Michael C. Bechner, Michael Place, Chris P. Churas, Louise Pape, Sally A. Leong, Rod Runnheim, Dan K. Forrest, Steve Goldstein, Miron Livny, and David C. Schwartz. Validation of rice genome sequence by optical mapping. *BMC Genomics*, 8 (1):278, 2007. ISSN 14712164. doi: 10.1186/1471-2164-8-278.
11. Thiago de Jesus Sousa, Doglas Parise, Rodrigo Profeta, Mariana Teixeira Dornelles Parise, Anne Cybelle Pinto Gomide, Rodrigo Bentos Kato, Felipe Luiz Pereira, Henrique Cesar Pereira Figueiredo, Rommel Ramos, Bertram Brenig, Artur Luiz da Costa da Silva, Preetam Ghosh, Debmalya Barh, Aristóteles Góes-Neto, and Vasco Azevedo. Re-sequencing and optical mapping reveals misassemblies and real inversions on Corynebacterium pseudotuberculosis genomes. *Scientific Reports*, 9(1):16387, dec 2019. ISSN 2045-2322. doi: 10.1038/s41598-019-52695-4.
12. Joshua A. Udall and R. Kelly Dawe. Is it ordered correctly? Validating genome assemblies by optical mapping. *Plant Cell*, 30(1):7–14, 2018. ISSN 1532298X. doi: 10.1105/tpc.17.00514.
13. Matthew Pendleton, Robert Sebra, Andy Wing Chun Pang, Ajay Ummat, Oscar Franzen, Tobias Rausch, Adrian M. Stütz, William Stedman, Thomas Anantharaman, Alex Hastie, Heng Dai, Markus Hsi Yang Fritz, Han Cao, Ariella Cohain, Gintaras Deikus, Russell E. Durrett, Scott C. Blanchard, Roger Altman, Chen Shan Chin, Yan Guo, Ellen E. Paxinos, Jan O. Korbel, Robert B. Darnell, W. Richard McCombie, Pui Yan Kwok, Christopher E. Mason, Eric E. Schadt, and Ali Bashir. Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nature Methods*, 12(8):780–786, 2015. ISSN 15487105. doi: 10.1038/nmeth.3454.
14. Yulia Mostovoy, Michal Levy-Sakin, Jessica Lam, Ernest T. Lam, Alex R. Hastie, Patrick Marks, Joyce Lee, Catherine Chu, Chin Lin, Zeljko Dzakula, Han Cao, Stephen A. Schlebusch, Kristina Giorda, Michael Schnall-Levin, Jeffrey D. Wall, and Pui Yan Kwok. A hybrid approach for de novo human genome sequence assembly and phasing. *Nature Methods*, 13(7):587–590, 2016. ISSN 15487105. doi: 10.1038/nmeth.3865.
15. Karyn Meltz Steinberg, Tina Graves Lindsay, Valerie Schneider, Mark Chaisson, Chad Tomlinson, John Huddleston, Patrick Minx, Milinn Kremitzki, Derek Albrecht, Vincent Magrini, Sean McGrath, Archana Raja, Carl Baker, Lana Harshman, LaDeana Hillier, Françoise Thibaud-Nissen, Nathan Bouk, Amy Ly, Chris Amemiya, Joyce Tang, Evan Eichler, Robert Fulton, Wesley Warren, Deanna Church, and Richard Wilson. High-Quality Assembly of an Individual of Yoruban Descent. *bioRxiv*, page 067447, 2016. doi: 10.1101/067447.
16. Shubha Vij, Heiner Kuhl, Inna S. Kuznetsova, Aleksey Komissarov, Andrey A. Yurchenko, Peter Van Heusden, Siddharth Singh, Natascha M. Thevasagayam, Sai Rama Sridatta Prakki, Kathiresan Purushothaman, Jolly M. Saju, Junhui Jiang, Stanley Kimbung Mbandi, Mario Jonas, Amy Hin Yan Tong, Sarah Mwangi, Doreen Lau, Si Yan Ngoh, Woei Chang Liew, Xueyan Shen, Lawrence S. Hon, James P. Drake, Matthew Boitano, Richard Hall, Chen-Shan Chin, Ramkumar Lachumanan, Jonas Korlach, Vladimir Trifonov, Marsel Kabilov, Alexey Tupikin, Darrell Green, Simon Moxon, Tyler Garvin, Fritz J. Sedlazeck, Gregory W. Vurture, Gopikrishna Gopalapillai, Vinaya Kumar Katneni, Tansyn H. Noble, Vinod Scaria, Sridhar Sivasubbu, Dean R. Jerry, Stephen J O'Brien, Michael C. Schatz, Tamás Dalmay, Stephen W. Turner, Si Lok, Alan Christoffels, and László Orbán. Chromosomal-Level Assembly of the Asian Seabass Genome Using Long Sequence Reads and Multi-layered Scaffolding. *PLoS genetics*, 12(4):e1005954, apr 2016. ISSN 1553-7404. doi: 10.1371/journal.pgen.1005954.
17. Andrea Minio, Mélanie Massonnet, Rosa Figueroa-Balderas, Alvaro Castro, and Dario Cantu. Diploid Genome Assembly of the Wine Grape Carménère. *G3: Genes, Genomes, Genetics*, 9(5):1331–1337, may 2019. ISSN 2160-1836. doi: 10.1534/g3.119.400030.
18. Mitchell R. Vollger, Philip C. Dishuck, Melanie Sorensen, Anne Marie E. Welch, Vy Dang, Max L. Dougherty, Tina A. Graves-Lindsay, Richard K. Wilson, Mark J.P. Chaisson, and Evan E. Eichler. Long-read sequence and assembly of segmental duplications. *Nature Methods*, 16(1):88–94, 2019. ISSN 15487105. doi: 10.1038/s41592-018-0236-3.
19. Aaron M. Wenger, Paul Peluso, William J. Rowell, Pi Chuan Chang, Richard J. Hall, Gregory T. Concepcion, Jana Ebler, Arkarachai Fungtammasan, Alexey Kolesnikov, Nathan D. Olson, Armin Töpfer, Michael Alonge, Medhat Mahmoud, Yufeng Qian, Chen Shan Chin, Adam M. Phillippy, Michael C. Schatz, Gene Myers, Mark A. DePristo, Jue Ruan, Tobias Marschall, Fritz J. Sedlazeck, Justin M. Zook, Heng Li, Sergey Koren, Andrew Carroll, David R. Rank, and Michael W. Hunkapiller. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology*, 37(10):1155–1162, 2019. ISSN 15461696. doi: 10.1038/s41587-019-0217-9.
20. Mélanie Massonnet, Noé Cochetel, Andrea Minio, Amanda M. Vondras, Jerry Lin, Aline Muyle, Jadran F. Garcia, Yongfeng Zhou, Massimo Delledonne, Summaira Riaz, Rosa Figueroa-Balderas, Brandon S. Gaut, and Dario Cantu. The genetic basis of sex determination in grapes. *Nature communications*, 11(1):2902, dec 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-16700-z.
21. Karen H. Miga, Sergey Koren, Arang Rhie, Mitchell R. Vollger, Ariel Gershman, Andrey Bzikadze, Shelise Brooks, Edmund Howe, David Porubsky, Glennis A. Logsdon, Valerie A. Schneider, Tamara Potapova, Jonathan Wood, William Chow, Joel Armstrong, Jeanne Fredrickson, Evgenia Pak, Kristof Tigyi, Milinn Kremitzki, Christopher Markovic, Valerie Maduro, Amalia Dutra, Gerard G. Bouffard, Alexander M. Chang, Nancy F. Hansen, Amy B. Wilfert, Françoise Thibaud-Nissen, Anthony D. Schmitt, Jon-Matthew Belton, Siddarth Selvaraj, Megan Y. Dennis, Daniela C. Soto, Ruta Sahasrabudhe, Gulhan Kaya, Josh Quick, Nicholas J. Loman, Nadine Holmes, Matthew Loose, Urvashi Surti, Rosa ana Risques, Tina A. Graves Lindsay, Robert Fulton, Ira Hall, Benedict Paten, Kerstin Howe, Winston Timp, Alice Young, James C. Mullikin, Pavel A. Pevzner, Jennifer L. Gerton, Beth A. Sullivan, Evan E. Eichler, and Adam M. Phillippy. Telomere-to-telomere assembly of a complete human X chromosome. *Nature*, jul 2020. ISSN 1476-4687. doi: 10.1038/s41586-020-2547-7.
22. Rachel M. Sherman and Steven L. Salzberg. Pan-genomics in the human genome era. *Nature Reviews Genetics*, 21(4):243–254, 2020. ISSN 14710064. doi: 10.1038/s41576-020-0210-7.
23. Chen Shan Chin, Paul Peluso, Fritz J. Sedlazeck, Maria Nattestad, Gregory T. Concepcion, Alicia Clum, Christopher Dunn, Ronan O'Malley, Rosa Figueroa-Balderas, Abraham Morales-Cruz, Grant R. Cramer, Massimo Delledonne, Chongyuan Luo, Joseph R. Ecker, Dario Cantu, David R. Rank, and Michael C. Schatz. Phased diploid genome assembly with single-molecule real-time sequencing. *Nature Methods*, 13(12):1050–1054, 2016. ISSN 15487105. doi: 10.1038/nmeth.4035.
24. Michael J. Roach, Simon A. Schmidt, and Anthony R. Borneman. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics*, 19(1): 460, dec 2018. ISSN 1471-2105. doi: 10.1186/s12859-018-2485-7.
25. Amanda M. Vondras, Andrea Minio, Barbara Blanco-Ulate, Rosa Figueroa-Balderas, Michael A. Penn, Yongfeng Zhou, Danelle Seymour, Zirou Ye, Dingren Liang, Lucero K. Espinoza, Michael M. Anderson, M. Andrew Walker, Brandon Gaut, and Dario Cantu. The genomic diversification of grapevine clones. *BMC genomics*, 20(1):972, 2019. ISSN 14712164. doi: 10.1186/s12864-019-6211-2.
26. Yongfeng Zhou, Andrea Minio, Mélanie Massonnet, Edwin Solares, Yuanda Lv, Tengizi Beridze, Dario Cantu, and Brandon S. Gaut. The population genetics of structural variants in grapevine domestication. *Nature Plants*, 5(9):965–979, 2019. ISSN 20550278. doi: 10.1038/s41477-019-0507-8.
27. Shengfeng Huang, Zelin Chen, Guangrui Huang, Ting Yu, Ping Yang, Jie Li, Yonggui Fu, Shaochun Yuan, Shangwu Chen, and Anlong Xu. HaploMerger: Reconstructing allelic relationships for polymorphic diploid genome assemblies. *Genome Research*, 22(8):1581–1588, 2012. ISSN 10889051. doi: 10.1101/gr.133652.111.
28. Leszek P. Pryszcz and Toni Gabaldón. Redundans: An assembly pipeline for highly heterozygous genomes. *Nucleic Acids Research*, 44(12):e113, 2016. ISSN 13624962. doi: 10.1093/nar/gkw294.
29. Jie Huang, Xinming Liang, Yuankai Xuan, Chunyu Geng, Yuxiang Li, Haorong Lu, Shoufang Qu, Xianglin Mei, Hongbo Chen, Ting Yu, Nan Sun, Junhua Rao, Jiahao Wang, Wenwei Zhang, Ying Chen, Sha Liao, Hui Jiang, Xin Liu, Zhaopeng Yang, Feng Mu, and Shangxian Gao. LR_Gapcloser: a tiling path-based gap closer that uses long reads to complete genome assembly. *GigaScience*, 6(5):1, jan 2017. ISSN 2047-217X. doi: 10/ggh6s3.
30. Han Ming Gan, Mun Hua Tan, Christopher M. Austin, Craig D. H. Sherman, Yen Ting Wong, Jan Strugnell, Mark Gervis, Luke McPherson, and Adam D. Miller. Best Foot Forward: Nanopore Long Reads, Hybrid Meta-Assembly, and Haplotig Purging Optimizes the First Genome Assembly for the Southern Hemisphere Blacklip Abalone (Haliotis rubra). *Frontiers in Genetics*, 10:889, sep 2019. ISSN 1664-8021. doi: 10.3389/fgene.2019.00889.
31. Dengfeng Guan, Dengfeng Guan, Shane A. McCarthy, Jonathan Wood, Kerstin Howe, Yadong Wang, Richard Durbin, and Richard Durbin. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics*, 36(9):2896–2898, 2020. ISSN 14602059. doi: 10.1093/bioinformatics/btaa025.
32. Xingtan Zhang, Ruoxi Wu, Yibin Wang, Jiaxin Yu, and Haibao Tang. Unzipping haplotypes in diploid and polyploid genomes. *Computational and Structural Biotechnology Journal*, 18: 66–72, 2020. ISSN 20010370. doi: 10.1016/j.csbj.2019.11.011.
33. Xiangchao Gan, Oliver Stegle, Jonas Behr, Joshua G. Steffen, Philipp Drewe, Katie L. Hildebrand, Rune Lyngsoe, Sebastian J. Schultheiss, Edward J. Osborne, Vipin T. Sreedharan, André Kahles, Regina Bohnert, Géraldine Jean, Paul Derwent, Paul Kersey, Eric J. Belfield, Nicholas P. Harberd, Eric Kemen, Christopher Toomajian, Paula X. Kover, Richard M. Clark, Gunnar Rätsch, and Richard Mott. Multiple reference genomes and transcriptomes for Arabidopsis thaliana. *Nature*, 477(7365):419–423, 2011. ISSN 00280836. doi: 10.1038/nature10414.
34. Gaik Tamazian, Pavel Dobrynin, Ksenia Krasheninnikova, Aleksey Komissarov, Klaus Peter Koepfli, and Stephen J. O'Brien. Chromosomer: A reference-based genome arrangement tool for producing draft chromosome sequences. *GigaScience*, 5(1):1–11, 2016. ISSN 2047217X. doi: 10.1186/s13742-016-0141-6.
35. Andrea Minio, Mélanie Massonnet, Rosa Figueroa-Balderas, Amanda M. Vondras, Barbara Blanco-Ulate, and Dario Cantu. Iso-Seq Allows Genome-Independent Transcriptome Profiling of Grape Berry Development. *G3 Genes|Genomes|Genetics*, 9(3):755–767, mar 2019. ISSN 2160-1836. doi: 10.1534/g3.118.201008.
36. Kun Tze Chen, Hsin Ting Shen, and Chin Lung Lu. Multi-CSAR: A multiple reference-based contig scaffolder using algebraic rearrangements. *BMC Systems Biology*, 12(Suppl 9):0–6, 2018. ISSN 17520509. doi: 10.1186/s12918-018-0654-y.
37. Huilong Du and Chengzhi Liang. Assembly of chromosome-scale contigs by efficiently resolving repetitive sequences with long reads. *Nature Communications*, 10(1):1–10, 2019. ISSN 20411723. doi: 10.1038/s41467-019-13355-3.

38. Haibao Tang, Xingtan Zhang, Chenyong Miao, Jisen Zhang, Ray Ming, James C. Schnable, Patrick S. Schnable, Eric Lyons, and Jianguo Lu. ALLMAPS: Robust scaffold ordering based on multiple maps. *Genome Biology*, 16(1):1–15, 2015. ISSN 1474760X. doi: 10.1186/s13059-014-0573-1.
39. Michael Alonge, Sebastian Soyk, Srividya Ramakrishnan, Xingang Wang, Sara Goodwin, Fritz J. Sedlazeck, Zachary B. Lippman, and Michael C. Schatz. RaGOO: Fast and accurate reference-guided scaffolding of draft genomes. *Genome Biology*, 20(1):1–17, 2019. ISSN 1474760X. doi: 10.1186/s13059-019-1829-6.
40. Alexey Gurevich, Vladislav Saveliev, Nikolay Vyahhi, and Glenn Tesler. QUAST: Quality assessment tool for genome assemblies. *Bioinformatics*, 29(8):1072–1075, apr 2013. ISSN 13674803. doi: 10.1093/bioinformatics/btt086.
41. Guillaume Marçais, Arthur L. Delcher, Adam M. Phillippy, Rachel Coston, Steven L. Salzberg, and Aleksey Zimin. MUMmer4: A fast and versatile genome alignment system. *PLoS Computational Biology*, 14(1):e1005944, jan 2018. ISSN 15537358. doi: 10.1371/journal.pcbi.1005944.
42. Heng Li. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18): 3094–3100, sep 2018. ISSN 14602059. doi: 10.1093/bioinformatics/bty191.
43. Jennafer A.P. Hamlin, Guilherme B. Dias, Casey M. Bergman, and Douda Bensasson. Phased diploid genome assemblies for three strains of Candida albicans from oak trees. *G3: Genes, Genomes, Genetics*, 9(11):3547–3554, 2019. ISSN 21601836. doi: 10.1534/g3.119.400486.
44. Anja Forche, P T Magee, B. B. Magee, and Georgiana May. Genome-wide single-nucleotide polymorphism map for Candida albicans. *Eukaryotic cell*, 3(3):705–14, jun 2004. ISSN 1535-9778. doi: 10.1128/EC.3.3.705-714.2004.
45. Tatjana Singer, Yiping Fan, Hur Song Chang, Tong Zhu, Samuel P. Hazen, and Steven P. Briggs. A high-resolution map of Arabidopsis recombinant inbred lines by whole-genome exon array hybridization. *PLoS Genetics*, 2(9):1352–1361, 2006. ISSN 15537390. doi: 10.1371/journal.pgen.0020144.
46. Sergey Koren, Arang Rhie, Brian P. Walenz, Alexander T. Dilthey, Derek M. Bickhart, Sarah B. Kingan, Stefan Hiendleder, John L. Williams, Timothy P.L. Smith, and Adam M. Phillippy. De novo assembly of haplotype-resolved genomes with trio binning. *Nature Biotechnology*, 2018. ISSN 15461696. doi: 10.1038/nbt.4277.
47. Wai Yee Low, Rick Tearle, Ruijie Liu, Sergey Koren, Arang Rhie, Derek M. Bickhart, Benjamin D. Rosen, Zev N. Kronenberg, Sarah B. Kingan, Elizabeth Tseng, Françoise Thibaud-Nissen, Fergal J. Martin, Konstantinos Billis, Jay Ghurye, Alex R. Hastie, Joyce Lee, Andy W.C. Pang, Michael P. Heaton, Adam M. Phillippy, Stefan Hiendleder, Timothy P.L. Smith, and John L. Williams. Haplotype-resolved genomes provide insights into structural variation and gene content in Angus and Brahman cattle. *Nature Communications*, 11(1), 2020. ISSN 20411723. doi: 10.1038/s41467-020-15848-y.
48. Amanda M. Vondras, Larry Lerno, Mélanie Massonnet, Andrea Minio, Adib Rowhani, Dingren Liang, Jadran Garcia, Daniela Quiroz, Rosa Figueroa-Balderas, Deborah A. Golino, Susan E. Ebeler, Maher Al Rwahnih, and Dario Cantu. Rootstock influences the effect of grapevine leafroll-associated viruses on berry development and metabolism via abscisic acid signalling. *Molecular Plant Pathology*, 22(8):984–1005, 2021. ISSN 13643703. doi: 10.1111/mpp.13077.
49. Noé Cochetel, Andrea Minio, Mélanie Massonnet, Amanda M Vondras, Rosa Figueroa-Balderas, and Dario Cantu. Diploid chromosome-scale assembly of the Muscadinia rotundifolia genome supports chromosome fusion and disease resistance gene expansion during Vitis and Muscadinia divergence . *G3 Genes|Genomes|Genetics*, 11(4), 2021. ISSN 2160-1836. doi: 10.1093/g3journal/jkab033.
50. Cheng Zou, Avinash Karn, Bruce Reisch, Allen Nguyen, Yongming Sun, Yun Bao, Michael S. Campbell, Deanna Church, Stephen Williams, Xia Xu, Craig A. Ledbetter, Sagar Patel, Anne Fennell, Jeffrey C. Glaubitz, Matthew Clark, Doreen Ware, Jason P. Londo, Qi Sun, and Lance Cadle-Davidson. Haplotyping the Vitis collinear core genome with rhAmpSeq improves marker transferability in a diverse genus. *Nature Communications*, 11(1):413, dec 2020. ISSN 2041-1723. doi: 10.1038/s41467-019-14280-1.
51. Benjamin D. Rosen, Derek M. Bickhart, Robert D. Schnabel, Sergey Koren, Christine G. Elsik, Elizabeth Tseng, Troy N. Rowan, Wai Y. Low, Aleksey Zimin, Christine Couldrey, Richard Hall, Wenli Li, Arang Rhie, Jay Ghurye, Stephanie D. McKay, Françoise Thibaud-Nissen, Jinna Hoffman, Brenda M. Murdoch, Warren M. Snelling, Tara G. McDaneld, John A. Hammond, John C. Schwartz, Wilson Nandolo, Darren E. Hagen, Christian Dreischer, Sebastian J. Schultheiss, Steven G. Schroeder, Adam M. Phillippy, John B. Cole, Curtis P Van Tassell, George Liu, Timothy P L Smith, and Juan F. Medrano. De novo assembly of the cattle reference genome with single-molecule sequencing. *GigaScience*, 9(3):1–9, mar 2020. ISSN 2047-217X. doi: 10.1093/gigascience/giaa021.
52. Riccardo Velasco, Andrey Zharkikh, Michela Troggio, Dustin A. Cartwright, Alessandro Cestaro, Dmitry Pruss, Massimo Pindo, Lisa M Fitzgerald, Silvia Vezzulli, Julia Reid, Giulia Malacarne, Diana Iliev, Giuseppina Coppola, Bryan Wardell, Diego Micheletti, Teresita Macalma, Marco Facci, Jeff T. Mitchell, Michele Perazzolli, Glenn Eldredge, Pamela Gatto, Rozan Oyzerski, Marco Moretto, Natalia Gutin, Marco Stefanini, Yang Chen, Cinzia Segala, Christine Davenport, Lorenzo Dematté, Amy Mraz, Juri Battilana, Keith Stormo, Fabrizio Costa, Quanzhou Tao, Azeddine Si-Ammour, Tim Harkins, Angie Lackey, Clotilde Perbost, Bruce Taillon, Alessandra Stella, Victor Solovyev, Jeffrey A. Fawcett, Lieven Sterck, Klaas Vandepoele, Stella M. Grando, Stefano Toppo, Claudio Moser, Jerry Lanchbury, Robert Bogden, Mark Skolnick, Vittorio Sgaramella, Satish K. Bhatnagar, Paolo Fontana, Alexander Gutin, Yves Van de Peer, Francesco Salamini, and Roberto Viola. A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. *PloS one*, 2(12): e1326, dec 2007. ISSN 1932-6203. doi: 10.1371/journal.pone.0001326.
53. Yongfeng Zhou, Aline Muyle, and Brandon S. Gaut. Evolutionary Genomics and the Domestication of Grapes. pages 39–55. Springer International Publishing, Cham, 2019. ISBN 978-3-030-18600-5 978-3-030-18601-2. doi: 10.1007/978-3-030-18601-2_3.
54. Olivier Jaillon, Jean Marc Aury, Benjamin Noel, Alberto Policriti, Christian Clepet, Alberto Casagrande, Nathalie Choisne, Sébastien Aubourg, Nicola Vitulo, Claire Jubin, Alessandro Vezzi, Fabrice Legeai, Philippe Hugueney, Corinne Dasilva, David Horner, Erica Mica, Delphine Jublot, Julie Poulain, Clémence Bruyère, Alain Billault, Béatrice Segurens, Michel Gouyvenoux, Edgardo Ugarte, Federica Cattonaro, Véronique Anthouard, Virginie Vico, Cristian Del Fabbro, Michaël Alaux, Gabriele Di Gaspero, Vincent Dumas, Nicoletta Felice, Sophie Paillard, Irena Juman, Marco Moroldo, Simone Scalabrin, Aurélie Canaguier, Isabelle Le Clainche, Giorgio Malacrida, Eléonore Durand, Graziano Pesole, Valérie Laucou, Philippe Chatelet, Didier Merdinoglu, Massimo Delledonne, Mario Pezzotti, Alain Lecharny, Claude Scarpelli, François Artiguenave, M. Enrico Pè, Giorgio Valle, Michele Morgante, Michel Caboche, Anne Françoise Adam-Blondon, Jean Weissenbach, Francis Quétier, and Patrick Wincker. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, 449(7161):463–467, sep 2007. ISSN 14764687. doi: 10.1038/nature06148.
55. A Canaguier, J Grimplet, G Di Gaspero, S Scalabrin, E Duchêne, N Choisne, N Mohellibi, C Guichard, S Rombauts, I Le Clainche, A Bérard, A Chauveau, R Bounon, C Rustenholz, M Morgante, M.-C. Le Paslier, D Brunel, and A.-F. Adam-Blondon. A new version of the grapevine reference genome assembly (12X.v2) and of its annotation (VCost.v3). *Genomics Data*, 14:56–62, dec 2017. ISSN 2213-5960. doi: https://doi.org/10.1016/j.gdata.2017.09.002.

Table 1. Assembly statistics

Genotype	Kingdom	Haploid Size	Chromo-somes	Techno-logy	Markers (per Mbp)	Input sequences		Results	HaploSplit	HaploFill
C. albicans	Fungi	14 Mbp	7 + R	PacBio [1]	116 (8.3) [2]	Primary	15.5 Mbp	Hap1	11.6 Mbp	12.9 Mbp
						Haplotigs	13.8 Mbp	Hap2	12.4 Mbp	13.7 Mbp
						Total	29.2 Mbp	Unpl.	5.2 Mbp	2.7 Mbp
A. thaliana	Plantae	119 Mbp	5	PacBio [3]	676 (5.7) [4]	Primary	140.0 Mbp	Hap1	109.0 Mbp	114.7 Mbp
						Haplotigs	104.9 Mbp	Hap2	106.6 Mbp	111.5 Mbp
						Total	245.0 Mbp	Unpl.	29.4 Mbp	19.0 Mbp
B. taurus	Animalia	2.62 Gbp (29+X) 2.49 Gbp (29 + Y)	29 + XY	PacBio [5]	46,325 (17.6) [6]	Primary	2.7 Gbp	Hap1	2.6 Gbp (29+X) 2.3 Gbp (29+Y)	2.6 Gbp (29+X) 2.5 Gbp (29+Y)
						Haplotigs	2.5 Gbp	Hap2		
						Total	5.2 Gbp	Unpl.	0.3 Gbp	0.2 Gbp
V. vinifera cv. Cabernet Franc	Plantae	480 Mbp	19	PacBio + Doveatil HiC [7]	1,661 (3.5) [8]	Primary	570.2 Mbp	Hap1	350.8 Mbp	455.6 Mbp
						Haplotigs	284.7 Mbp	Hap2	263.4 Mbp	410.9 Mbp
						Total	854.9 Mbp	Unpl.	239.9 Mbp	47.1 Mbp
M. rotundifolia	Plantae	483 Mbp	20	PacBio + BioNano [9]	1,661 (3.5) [10]	Primary	459.5 Mbp [11]	Hap1	374.3 Mbp	400.5 Mbp
						Haplotigs	364.8 Mbp [12]	Hap2	338.9 Mbp	370.0 Mbp
						Total	896.0 Mbp	Unpl.	165.5 Mbp	63.0 Mbp

1. FalconUnzip Hamlinet al. (2019)
2. Forche et al. (2004)
3. FalconUnzip, Chinet al. (2016)
4. Singer et al. (2006)
5. FalconUnzip, Koren et al. (2018)
6. Low et al. (2020) using the Integrated Bovine Map of sex chromosome (ver. Btau_4.0, https://www.hgsc.bcm.edu/other-mammals/bovine-genome-project)
7. FalconUnzip + SSPACE + HiRise, Vondras et al. (2021)
8. Zou et al. (2020)
9. FalconUnzip + Hybrid Scaffolder, Cochetel et al. (2021)
10. Zou et al. (2020)
11. Reported for FalconUnzip assembly as haplotype separation is lost during Hybrid Scaffolding
12. Reported for FalconUnzip assembly as haplotype separation is lost during Hybrid Scaffolding